# On (Emergent) Systematic Generalisation and Compositionality in Visual Referential Games with Straight-Through Gumbel-Softmax Estimator

**Kevin Denamganaï and James Alfred Walker**
Department of Computer Science
University of York
York, UK
kyd500@york.ac.uk, james.walker@york.ac.uk

## Abstract

The drivers of compositionality in artificial languages that emerge when two (or more) agents play a non-visual referential game has been previously investigated using approaches based on the REINFORCE algorithm and the (Neural) Iterated Learning Model. Following the more recent introduction of the *Straight-Through Gumbel-Softmax* (ST-GS) approach, this paper investigates to what extent the drivers of compositionality identified so far in the field apply in the ST-GS context and to what extent do they translate into (emergent) systematic generalisation abilities, when playing a visual referential game. Compositionality and the generalisation abilities of the emergent languages are assessed using topographic similarity and zero-shot compositional tests.

Firstly, we provide evidence that the test-train split strategy significantly impacts the zero-shot compositional tests when dealing with visual stimuli, whilst it does not when dealing with symbolic ones.

Secondly, empirical evidence shows that using the ST-GS approach with small batch sizes and an overcomplete communication channel improves compositionality in the emerging languages. Nevertheless, while shown robust with symbolic stimuli, the effect of the batch size is not so clear-cut when dealing with visual stimuli. Our results also show that not all overcomplete communication channels are created equal. Indeed, while increasing the maximum sentence length is found to be beneficial to further both compositionality and generalisation abilities, increasing the vocabulary size is found detrimental.

Finally, a lack of correlation between the language compositionality at training-time and the agents' generalisation abilities is observed in the context of discriminative referential games with visual stimuli. This is similar to previous observations in the field using the generative variant with symbolic stimuli.

## 1 Introduction

In recent years, research into language emergence and grounding have received increased attention. The former of which raises the question of how to make artificial languages emerge with similar properties to natural languages, or at least 'natural-like' protolanguages, exhibiting compositionality as the primarily-targeted property[4, 21, 45, 60]. Indeed, languages' compositionality has been shown to further the learnability of said languages [34, 64, 9, 45] and promises to increase the generalisation ability of the artificial agent that would be able to wield them. For instance, it has been found to be instrumental in producing learned representations that generalise, when measured

4th Emergent Communication Workshop (NeurIPS 2020), Vancouver, Canada.

in terms of the data-efficiency of subsequent transfer and/or curriculum learning [25, 54, 55, 32]. Nevertheless, the ability of neural networks to generalise in a systematic fashion has been called into question [42, 47, 46, 3], and investigated towards finding necessary conditions and/or paradigms that favour the emergence of systematicity [26, 63, 38, 41, 62]. In this paper, similarly to Hill et al. [26] investigating natural language grounding in the context of embodied agents, we take a closer look at the conditions that further the emergence of compositionality in artificial languages in the context of referential games.

**Straight-Through Gumbel-Softmax and Visual Stimuli.** Although it has been shown that emerging languages are far from being 'natural'-like [39, 12, 13], there are some successful cases demonstrating the emergence of compositional languages and learned representations (e.g. Kottur et al. [39], Lazaridou et al. [44], Choi et al. [15], Bogin et al. [7], Guo et al. [21], Korbak et al. [37], Chaabouni et al. [14]), relative to a given standard of compositionality. This paper focuses exclusively on the *Straight-Through Gumbel-Softmax* (ST-GS) approach proposed by Havrylov and Titov [23], as it supposedly allows a richer signal towards solving the credit assignment problem that language emergence poses since the gradient can be backpropagated from the listener agent to the speaker agent, while, in comparison, it cannot be backpropagated when using more commonly adopted approaches based on REINFORCE-like algorithms [69].

**Symbolic vs Visual Stimuli.** The main works in language emergence and grounding focus on symbolic/one-hot-encoded stimuli (e.g. [39, 14]) whereas the ultimate goal of the field is related to grounding in similar modalities enjoyed by human beings, and primarily sight. Therefore, in order to take one more step in this direction, in this paper, we investigate primarily visual/pixel-based stimuli as well as whether results found in the context of symbolic stimuli translates into the context of visual stimuli.

**Compositionality & Generalisation.** As a concept, compositionality has been the focus of many definition attempts. For instance, it can be defined as "the algebraic capacity to understand and produce novel combinations from known components"(Loula et al. [47] referring to Montague [53]) or as the property that sees "the meaning of a complex expression [a]s a function of the meaning of its immediate syntactic parts and the way in which they are combined" [40]. Although difficult to define, the commmunity seem to agree on the fact that it would enable learning agents to exhibit systematic generalisation abilities (also referred to as combinatorial generalisation [5]). Some of the ambiguities that come with those loose definitions start to be better understood and explained, as in the work of Hupkes et al. [29]. In this paper, we will refer to compositionality as "the ability to entertain a given thought implies the ability to entertain thoughts with semantically related contents"[19], and thus use it interchangeably with systematicity, following the classification made by Hupkes et al. [29].

Compositionality, as a property of languages, can be difficult to measure. Like Keresztury and Bruni [33], we acknowledge that it ought to be measured more quantitatively than doing qualitative eye-balling of the artificial agents' utterances. Therefore, this paper aims to provide a more complete picture of the phenomena involving the ST-GS approach using two metrics: *topographic similarity* [11], which is acknowledged by the research community as the main quantitative metric for compositionality [44, 21, 63, 14, 60], and *zero-shot compositional tests*, in which a set of stimuli composed of specific attributes are held-out from the training set, while making sure that the agents are still familiarising themselves with the specific attributes in different contexts/combinations to the held-out ones. The use of these two metrics has been shown to be critical by the concurrent work of Chaabouni et al. [14], which shows that learning agents are able to generalise in a referential game in spite of their utterances not being compositional, when measured with *topographic similarity*.

**Contributions.** Firstly, this paper questions whether the train-test split strategy matters when using *zero-shot compositional tests*. Out of the two strategies tested (see Section 3.1), it was found that the choice significantly impacts the metric when dealing with visual stimuli, whilst it does not when dealing with stimuli that are symbolic/one-hot-encoded (see Section 4.1). Appendix E further investigates this impact using the lens of the concept of generalisation difficulty proposed by Chollet [16].

Secondly, following the definition of compositionality/systematicity and the use of both topographic similarity and zero-shot compositional test metrics, this paper provides a quantitative report on the extent with which the ST-GS relaxation is a viable approach to make compositional languages emerge in a (discriminative) referential game. We start by identifying the batch size hyperparameter as

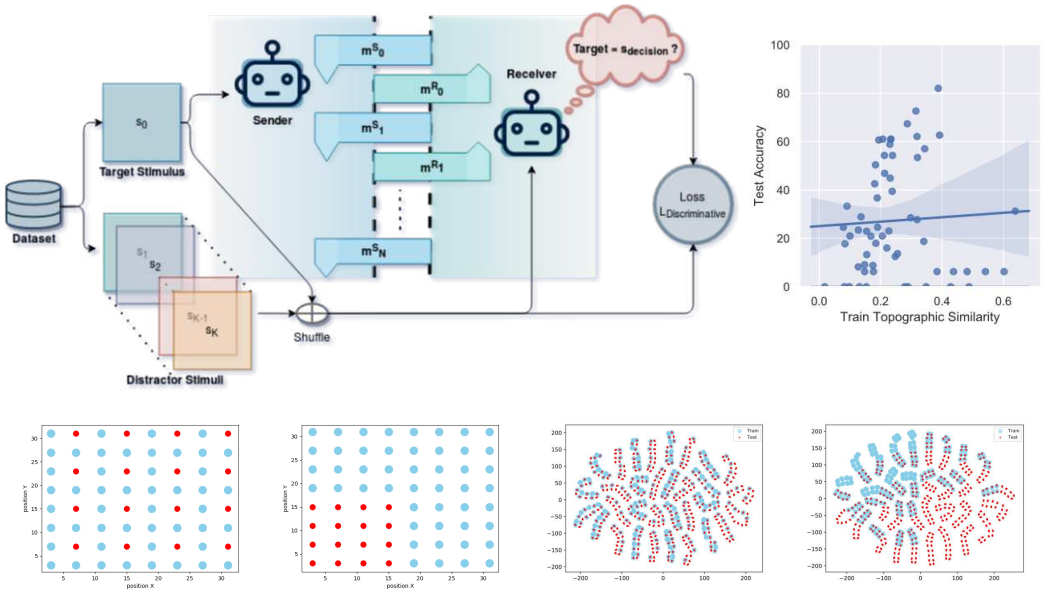


Figure 1: **Top left:** Illustration of a *partially-observable* 2*-players/L-signal/*$N$*-round/uniformly-distributed-distractors referential game*. **Top right:** Test-time accuracy with respect to the training-time topographic similarity across the different settings, along with a linear regression model fit for comparison. **Bottom:** $2D$ projection of the symbolic 2-attribute benchmarks (left) and T-SNE representations of the symbolic 3-attribute benchmarks (right), for both *interpolation* and *extrapolation* tasks (from left to right). Blue dots correspond to training examples (their vicinity is emphasised by drawing them larger in the t-SNE), while red dots correspond to testing examples.

a potential driver of compositionality (see Section 4.2.1). Whilst our experiments depict it as a robust driver of compositionality in the context of symbolic stimuli, its impact on compositionality is less significant when dealing with visual stimuli. Subsequently, we investigate the impact on compositionality of both the level of structure in the observed stimuli and the capacity of the communication channel (see Section 4.2.2). Our experiments show that, using the ST-GS approach, they both correlate positively with the agents generalisation abilities. More specifically, we found that increasing the capacity of the communication channel by means of increasing the maximum sentence length correlates positively with both the compositionality of the language and the generalisation abilities of the learning agents, but increasing it by means of increasing the vocabulary size is negatively correlated with our two metrics. Appendix B provides theoretical insights on these phenomenon by comparing the ST-GS approach to the (Neural) *Iterated Learning Model* (ILM) [35, 60].

Finally, similarly to Hupkes et al. [29], this paper investigates under what circumstances a model can be called compositional, and more precisely when does a learning agent is said to have systematicity. We ask whether a learning agent has systematicity/compositional abilities when the "utterances in the arena of use" [30] that it produces are compositional? While there seems to be such an implicit assumption in the research community , the results presented in Section 4.2.3 disagree with this assumption. This is similar to the results observed in the concurrent and related work of Chaabouni et al. [14], in which symbolic/one-hot-encoded stimuli and a generative referential game are used, instead of visual stimuli in a discriminative framing as in this paper (see Denamganaï and Walker [18] for further details on the different variants of referential games).

## 2 Preliminaries

### 2.1 Visual Referential Games

Referential/Language games emphasise the functionality of languages, namely, the ability to efficiently communicate and coordinate between agents. Following the nomenclature proposed in

Denamganaï and Walker [18], we will focus primarily on a *partially-observable* 2*-players/L-signal/*0*-round/uniformly-distributed-distractors* variant illustrated in Figure 1. As it is a discriminative referential game, we specify the number of distractor stimuli $K$. At training-time, $K = 47$, and at testing-time, $K = 63$, which corresponds to the maximum number of distractors that can be used across the different benchmarks, which are described in Section 3. For a fair comparison of the measure of accuracy between benchmarks, the values of $K$ are fixed. Details about the agent's architecture can be found in Appendix C and in our code[1]. Having described the setup, the following section provides details of the kind of stimuli and the train/test set split that is used to evaluate the generalisation abilities of the tested learning agents, whilst emphasising where this study lies in the spectrum of generalisation evaluation.

### 2.2 Generalisation Abilities

While many studies have investigated the generalisation abilities of neural networks in the context of linguistic tasks (e.g. Kottur et al. [39], Lake and Baroni [42], Resnick et al. [61]), their assumed definition to the act of generalisation do not always coincide [29]. Following the work of Lake and Baroni [42] showing that recurrent neural networks (RNNs) fail to generalize systematically but are very successful at generalizing when enough supporting evidence is provided. It is worth emphasizing again, like many previous authors [3, 29, 62, 16], that it is critical to clearly state what kind of generalisation abilities this study aims to evaluate. The remainder of this section proposes an in-depth examination of the range of possible generalisation abilities and suggests where this paper lies within those identified.

While some studies test for what is referred to as systematic generalisation abilities [39, 42, 46, 38, 3, 62], others seem to test for (vanilla) generalisation abilities [2, 29, 14]. This work differentiates between the two in terms of the size of the pool of supporting evidence/samples, which the learning system is trained on. Note that pool size is assumed to vary the difficulty of the task. Thus, (hard) systematic generalisation being on the side of the spectrum where the pool size is the smallest, and (easy) (vanilla) generalisation being on the other side where it is the greatest. In this work, we focus on systematic generalisation, rather than vanilla, and we adopt Bahdanau et al. [3]'s definition of systematic generalisation abilities (i.e. a model that has systematic generalisation abilities "should be able to reason about all possible object combinations despite being trained on a very small subset of them"). This definition is a continuation of our assumed definition for compositional abilities, particularly "the ability to entertain a given thought implies the ability to entertain thoughts with semantically related contents"[19]. In other words, the kind of generalisation abilities we propose to test for could be referred as systematic combinatorial/compositional generalisation abilities.

### 2.3 Emergent Systematicity & Prior Knowledge

The work of Hill et al. [26] is particularly relevant to our current context, as it evaluates a specific kind of systematic generalisation referred to as "emergent systematicity" ( Hill et al. [26] citing McClelland et al. [52]), "because the architecture of our agent does not include components that are explicitly engineered to promote systematicity" (i.e. systematic generalisation abilities). The concurrent work of Chaabouni et al. [14] also tackles "emergent" systematicity/systematic generalisation. The learning agents will be equipped with CNNs and RNNs (see details in Appendix C), giving them some form of spatial/translational invariance prior, hierarchical/distributed visual feature prior, and sequence-computation-enabling prior.

Chollet [16] emphasizes that **priors**, **experience**, and **generalisation difficulty** must be controlled in order to reliably evaluate broad generalisation abilities. Acknowledging the context of emergent systematicity in which this study lies can therefore be understood as a step towards controlling the prior knowledge baked into deep learning agents when evaluating their generalisation abilities. While the matter of controlling the generalisation difficulty is difficult to address in the current context, it is broached upon in Appendix E. The matter of how to control each for the amount of experience available to the agent is discussed in Section 3.

[1]Our code is released at: `https://github.com/Near32/ReferentialGym/tree/master/zoo/referential-games%2Bst-gs`

## 3 Evaluation Methodology

In the following experiments, learning agents will observe both symbolic and visual stimuli from particular train/test splits of the dSprites dataset [51, 24]. Symbolic representations are obtained by one-hot-encoding the latent representations that the dataset provides. Originally employed as a benchmark for disentangled representation learning, the dSprites dataset consists of visual representations for combinations of values along some generative factors/attributes/latent axes. There are $32$ possible values on each position axis, X and Y, $40$ possible values on the Orientation axis, $6$ possible values on the Scale axis, and $3$ possible values on the Shape axis. Note that in our experiments the value on the Shape attribute is always fixed to be the *heart* shape in order to remove any orientation ambiguities (e.g. the square shape has four symmetries that prevents visual differentiation from, for instance, a rotation of $90$ degrees and that of $180$ degrees). The choice of the dSprites dataset is motivated by the availability of the generative factors (that are used as symbolic stimuli in their one-hot-encoded form), which permits the computation of topographic similarity [11] (following a similar approach to Lazaridou et al. [44]) to assess the degree of compositionality in the emerging languages, as well as doing so separately from assessing generalisation abilities. We emphasise again how critical the separation of the two measures is given the related and concurrent work of Chaabouni et al. [14] showing that learning agents are able to generalise in a referential game in spite of their utterances not being compositional.

The amount of prior experience will be controlled in two ways. Firstly, by limiting the diversity in the training set, thus constraining our study to very small data set sizes. Secondly, by limiting the number of training epochs to fit to a sample budget of $480,000$ training samples, across the different settings. The number of gradient steps is the sample budget divided by the batch size assumed, which is detailed for each experiments in Section 4.

### 3.1 Train-Test Split Strategy

In the first experiment, we question whether the train-test split strategy matters when testing zero-shot compositional abilities. We arbitrarily propose to look at two simple train-test split strategies leading up to two different sets of tasks that we will refer to as: the interpolation tasks, $T_{inter}$, and the extrapolation tasks, $T_{extra}$. Considering that each stimulus can be described as a set of values taken from each attribute/generative factor/latent axis of the dSprites dataset, details of how the train-test splits are performed in each task can be described as follows. In the interpolation tasks, $T_{inter}$, the test sets consist of defining testing-purpose (and not test-only) values alternating with other values, on each attribute axis. In the extrapolation tasks, $T_{extra}$, the test sets consist of defining the testing-purpose (and not test-only) values as the first values (following the ordering of the original dataset) on each attribute axis. We highlight already, and will detail further below, that testing-purpose values are encountered by the agent both at training-time and testing-time, just not in combinations with all other possible values on the other latent axises of which a stimuli is composed.

Independently of the task, half of the possible values for each attribute axis are defined as testing-purpose values. The choice of using half of the possible values is motivated by the results of Bahdanau et al. [3] when evaluating for emergent systematicity, i.e. a CNN+LSTM architecture similar to that of ours here, and showing that such a benchmark is already challenging enough. Then, any sampled stimulus that combines **two or more** testing-purpose values for its different attributes is automatically retained for the test set. In other words, the testing-purpose values on each latent axis are presented to the agent at training-time while in combinations with sole non-testing-purpose values on the other latent axises, in order for the agent to familiarise itself with all the possible values on each latent axis while remaining unaware of a subset of all the combinations possible. For an intuitive understanding, Figure 1 renders $2D$ projection of the symbolic 2-attributes tasks, $T^2_{inter}$ and $T^2_{extra}$, and the results of performing t-SNE [48] on the symbolic 3-attributes tasks, $T^3_{inter}$ and $T^3_{extra}$, thus highlighting their systematic and topological differences, and revealing the motivation for their naming as interpolation and extrapolation tasks. In a similar manner to Russin et al. [62], we acknowledge the intuitive difference of difficulty between the two flavours in which generalisation occurs [50]. These are: *interpolation*, where "the train and test sets are independent and identically distributed (i.i.d)", and *extrapolation*, where the tested system is required to make "an inferential leap about the entire structure of part of the distribution that they have not seen", and so the test set is out-of-domain (o.o.d.) with regards to the train set. When testing for combinatorial generalisation abilities, following Chollet [16]'s framework, it is hypothesised (and further detail in Appendix E)

that tasks involving compositional generalisation by *interpolation*, $T_{inter}$, have *zero generalisation difficulty*, while tasks involving compositional generalisation by *extrapolation*, $T_{extra}$, bear *some generalisation difficulty*. We emphasise again that the naming convention adopted here for the tasks is solely motivated by the intuition provided by figure 1, but the actual nature of the tasks are unknown, since we are dealing with stimuli represented as combinations of attributes. They cannot and should not be confounded with regular interpolation and extrapolation tasks, as far as we know.

### 3.2 Control for the Level of Structure in the Meaning Space

In the subsequent experiments, we will vary the level of structure in the meaning space by experimenting with subsets of the dSprites dataset, emphasizing either: 2 attributes (position on the X-axis and Y-axis, yielding $48$ training examples and $16$ testing examples), 3 attributes (similar to 2 attributes plus Orientation, yielding $256$ examples in each train/test set), or $4$ attributes (similar to 3 attributes plus Scale, yielding $960$ training examples and $2112$ testing examples). We leave it to future works to vary the level of structure by varying the number of possible values for each attribute.

The employed subsets subsample the original dSprites dataset, in order to reduce the number of samples available to the learning agents. This subsampling aims to control for the amount of prior experience, and thus design a test where the learning agents are not presented with oversized data sets from which to slowly build evidence [42, 47, 46]. With the exception of the Scale attribute that only originally contains $6$ possible values (and thus remains unchanged), each other attribute contains $8$ possible values that are sampled from the original data to be evenly spaced out (i.e. out of the $32$ possibles values on the X and Y attributes, we sample every $4$ values; out of the $40$ possibles values on the Orientation attribute, we sample every $5$ values).

### 3.3 Control for the Capacity of the Communication Channel

Two different cases are considered in the following experiments. The *complete* case, where there are exactly $8$ ungrounded symbols in the vocabulary $V$, plus a ninth grounded symbol, in order to account for the *end of sentence* semantic, thus $|V| = 9$. The maximum sentence length $L$ is always equal to the number of attributes in the subset on which the experiment takes place. On the other hand, the *overcomplete* case consists of a vocabulary of size $|V| = 100$, and maximum sentence length $L = 20$.

## 4 Experiments

### 4.1 Experiment 1: Impact of Train-Test Split Strategy

We firstly question whether the train-test split strategy matters when using *zero-shot compositional tests*. We investigate in the contexts of 3 attributes with symbolic and visual stimuli. Figure 2(top-left) shows that, independently of the capacity of the communication channel, the choice of the train-test split significantly impacts the metric when dealing with visual stimuli, as a wide gap of performance can be seen between the two tasks, whilst it does not when dealing with symbolic stimuli.

### 4.2 Experiment 2: Drivers of Compositionality

#### 4.2.1 Effect of the Batch Size

In this section, we investigate the hypothesis that the batch size modulates a transmission bottleneck effect in the context of the ST-GS algorithm, similarly to the effect highlighted in the ILM where the severity of the bottleneck (or its inverse, the coverage) is positively (negatively) correlated with the likelihood of highly-compositional language emergence. The soundness of this hypothesis is detailed in Appendix B.3. We firstly report on the the 3-attributes setting, with $10$ random seeds. Results obtained with symbolic and visual stimuli are presented, respectively, in figure 2 (right) and figure 2 (right). Once again, the dynamics are different, depending on the nature of the stimuli. While the symbolic context shows great support to the identification of the batch size as a robust driver of compositionality, its impact is less clear-cut when dealing with visual stimuli, as most of the p-values in our Kolmogorov-Smirnov (KS) two-samples tests are high. In this visual setting, we suspect that the batch normalization layers in the learning agents' CNNs starts playing a greater regularisation

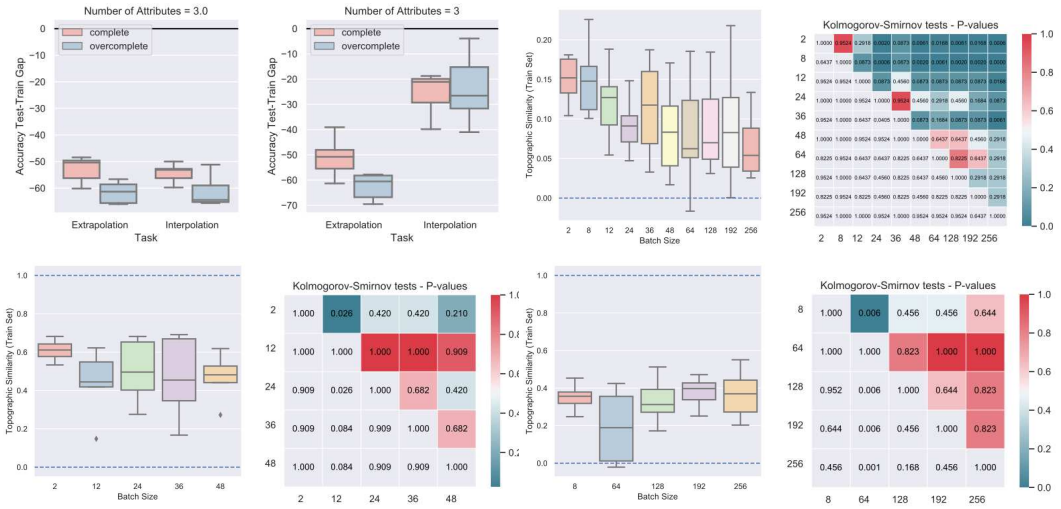


Figure 2: **Top left**: Distributions of the test-train gaps in accuracy measured on the 3-attributes benchmarks with, ordered from left to right, symbolic stimuli and visual stimuli. **Top right:** Distributions of training-time topographic similarity measured on the 3-attributes benchmarks with symbolic stimuli.The heatmaps illustrates the p-values of the KS two-samples tests between batch sizes with the alternative hypothesis being that the row-id distribution is 'greater' than the column-id one. **Bottom**: Distributions of training-time topographic similarity measured, with respect to different batch sizes and levels of structure, on the 2- (**left**) and 3-attributes (**right**) benchmarks. The heatmaps illustrate the p-values of the KS two-samples tests between batch sizes with the alternative hypothesis being that the row-id distribution is 'greater' than the column-id one.

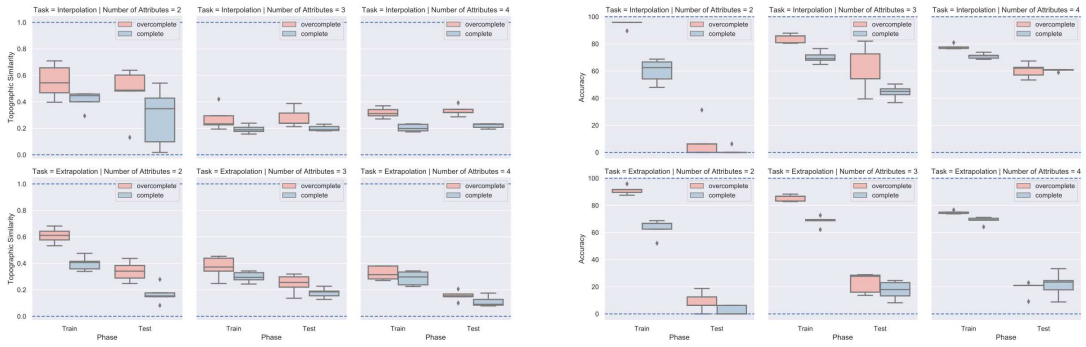


Figure 3: Distributions of the topographic similarity (left) and accuracy (right), across the different benchmarks, and the different communication channels.

role as, in this 3-attributes setting, even the smallest coverage value already entails to relatively high batch sizes.

Knowing that the level of structure in the observed meaning space has also been highlighted as a driver of compositionality in the context of the ILM, we further investigate the context of visual stimuli by benchmarking against the 2-attributes setting, this time with 5 random seeds for each batch size. Figure 2 (left) illustrates the results. Although not always statistically significant, in this level of structure, the likelihood of seeing highly-compositional languages seems to negatively correlate with the severity of the transmission bottleneck.

Comparing the results across the different levels of structure, it is surprising to see that, the topographic similarity measured is significantly lower when the level of structure is higher ($p \approx 3 \times 10^{-4}$ on a KS test between benchmarks, at the lowest $\sim 4\%$ coverage). Thus, on the contrary to the phenomenon observed in the ILM context, the ST-GS approach does not necessarily further higher compositionality in higher levels of structure. Finally, independently of the level of structure, the

Table 1: Results of Spearman rank-order correlation tests between the measured topographic similarity of the emerging languages (**top**), or the accuracy on the zero-shot compositional tests (**bottom**), and, respectively, the maximum sentence length $L \in \{3, 6, 9, 20\}$ (for $V \in \{9, 20\}$), or the vocabulary size $V \in \{9, 20, 50, 100\}$ (for $L \in \{10, 20\}$). 5 seeds in each communication channel capacity.

| Task | Phase | $V = 9$ | $V = 20$ | $L = 10$ | $L = 20$ |
|---|---|---|---|---|---|
| $T^3_{inter}$ | Train | $0.65$ ($p \approx 0.002$) | $0.46$ ($p \approx 0.043$) | $-0.18$ ($p \approx 0.452$) | $-0.35$ ($p \approx 0.132$) |
| | Test | $0.52$ ($p \approx 0.019$) | $0.09$ ($p \approx 0.721$) | $-0.07$ ($p \approx 0.770$) | $-0.21$ ($p \approx 0.376$) |
| $T^3_{extra}$ | Train | $0.63$ ($p \approx 0.003$) | $0.50$ ($p \approx 0.023$) | $-0.64$ ($p \approx 0.003$) | $0.0$ ($p \approx 1.0$) |
| | Test | $0.31$ ($p \approx 0.183$) | $0.47$ ($p \approx 0.039$) | $-0.47$ ($p \approx 0.039$) | $0.13$ ($p \approx 0.580$) |
| $T^3_{inter}$ | Train | $0.87$ ($p \approx 10^{-6}$) | $0.73$ ($p \approx 10^{-4}$) | $-0.02$ ($p \approx 0.948$) | $-0.56$ ($p \approx 0.010$) |
| | Test | $0.77$ ($p \approx 10^{-4}$) | $0.87$ ($p \approx 10^{-6}$) | $0.51$ ($p \approx 0.021$) | $-0.13$ ($p \approx 0.579$) |
| $T^3_{extra}$ | Train | $0.76$ ($p \approx 10^{-4}$) | $0.49$ ($p \approx 0.029$) | $-0.43$ ($p \approx 0.060$) | $-0.05$ ($p \approx 0.845$) |
| | Test | $-0.23$ ($p \approx 0.323$) | $0.26$ ($p \approx 0.261$) | $0.18$ ($p \approx 0.442$) | $0.13$ ($p \approx 0.590$) |

lowest coverage always yields relatively high topographic similarity languages, with relatively low variance in the distributions. Therefore, we subsequently fix the batch size to accommodate those lowest coverage values around $4\%$ (i.e. $2$ in the 2-attributes, $8$ in the 3-attributes, and $40$ in the 4-attributes benchmarks).

#### 4.2.2 Effect of Input/Meaning Space Structure and Communication Channel Capacity

We investigate the impact of both the level of structure in the observed meaning space and the capacity of the communication channel.Figure 3 (left) illustrates the impact in terms of topographic similarity. The most striking result is that, independently of the task, the overcomplete communication channel promotes significantly higher compositionality in the emerging languages ($p \approx 9 \times 10^{-4}$ and $p \approx 0.037$ for one-sided KS tests with the alternative hypothesis being that the overcomplete channel yields greater topographic similarity than the complete one, respectively, at testing-time and training-time, for the interpolation task ; $p \approx 0.0137$ and $p \approx 0.037$, similarly, for the extrapolation task). As the level of structure increases, the compositionality of the emerging languages using the ST-GS approach decreases, which is contrary to what can be observed in the context of the ILM.

In terms of the generalisation abilities of the learning agents, we report the results in figure 3 (right). The ST-GS approach seems to significantly favour overcomplete communication channels over complete ones. Independently of the task, overcomplete communication channels significantly yield greater data-efficiency at training time ($p \approx 1 \times 10^{-7}$ and $p \approx 6 \times 10^{-9}$ for one-sided KS tests with the alternative hypothesis being that the overcomplete channel yields greater accuracy than the complete one, on the interpolation and extrapolation tasks, respectively). The effect appears to diminish as the level of structure in the observed meaning space increases. We observe a decrease of the test-train generalisation gap as the amount of structure in the meaning space increases. As this increase of structure implies an increase in size of the input/meaning space, our results provides evidence, this time in the context of visual stimuli, of the effect observed by Chaabouni et al. [14] with symbolic stimuli that "generalisation emerges 'naturally' if the input space is large"[14].

We further investigate the impact of varying the capacity of the communication channel when the input space structure is kept fixed, using the 3-attribute benchmarks, $T^3_{extra}$ and $T^3_{inter}$. In these experiments, we vary the maximum sentence length $L$, while keeping the vocabulary size $V$ fixed, and vice versa (see Appendix A for the graphs). The results of conducting Spearman rank-order correlation tests with topographic similarity or with accuracy of zero-shot compositional tests are described in table 1. Independently of the task, increasing $L$ translates mainly in an increase in topographic similarity, at training-time, as well as in an increase in systematicity, overall. On the other hand, increasing $V$ is marginally detrimental.

These results are all the more surprising given that, in the context of symbolic stimuli, REINFORCE-like algorithms have been shown to require constrained communication channels to yield highly compositional languages [39], and that its capacity negatively correlates with compositionality [14]. Here, using the ST-GS approach with visual stimuli, the emergence of both compositionality and systematicity is furthered by overcomplete channels with high maximum sentence length.

#### 4.2.3 Lack of Correlation between Generalisation and Compositionality at Training-time

We now investigate the validity of the common assumptions that observing compositionality at training-time, as reported by topographic similarity, correlates with the systematicity of the learning agents. Figure 1 (right) shows the test-time accuracy with respect to the training-time topographic similarity across the different settings (see Appendix D for a break down by setting). The results of a Spearman-rho test yields a correlation coefficient of $0.136$ and $p \approx 0.302$. Not only the coefficient is arguing towards a lack of correlation, the large p-value also prevents us from rejecting the non-correlation hypothesis, thus showing that it is incorrect to expect the training-time topographic similarity to predict any systematicity of the learning agents in the context of visual stimuli.

## 5 Conclusion

In this paper, we have shown that the choice of train-test split in zero-shot compositional tests significantly impacts the metric when dealing with visual stimuli, whilst it does not when dealing with symbolic ones. Then, we provided a quantitative report on the extent with which the ST-GS relaxation is a viable approach to make compositional languages emerge in a (discriminative) referential game. Fistly, while the batch size hyperparameter is identified as a critical driver of compositionality, its impact is significant in the context of symbolic stimuli but less so clearly understood when dealing with visual stimuli. Secondly, overcomplete communication channels with large maximum sentence lengths are found to be beneficial to further both compositionality and systematicity while those with large vocabulary sizes is found detrimental. Our paper has highlighted many ways in which results about language emergence and grounding in the case of symbolic stimuli do not translate equally in the context of visual stimuli, thus raising the need for subsequent work in the visual context as vision is one of the main modalities of concerns for the grounding of languages that would be spoken by artificial agents living alongside and cooperating with human beings.

## Broader Impact

This work consists solely of simulations, thus evacuating some of the ethical concerns, as well as the concerns with regards to the consequences of failure of the system presented. With regards to the ethical aspects related to its inclusion in the field of Artificial Intelligence, we argue that our work aims to have positive outcomes on the development of human-machine interfaces, albeit being not yet mature enough to aim for this goal. The current state of our work does not allow us to extrapolate towards negative outcomes.

This work should benefit the research community of language emergence and grounding, in its current state.

## Acknowledgments and Disclosure of Funding

This work was supported by the EPSRC Centre for Doctoral Training in Intelligent Games & Games Intelligence (IGGI) [EP/L015846/1].

We would like to thank Dr Sondess Missaoui and the anonymous reviewers for their very helpful and constructive feedback on the draft of this paper.

We gratefully acknowledge the use of Python[66], IPython[59], SciPy[67], Scikit-learn[58], Scikit-image[65], NumPy[22], Pandas[68, 56], OpenCV[8], PyTorch[57], TensorboardX[28], Tensorboard from the Tensorflow ecosystem[1], without which this work would not be possible.

## References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and

X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

[2] J. Andreas. Measuring Compositionality in Representation Learning. feb 2019. URL http://arxiv.org/abs/1902.07181.

[3] D. Bahdanau, S. Murty, M. Noukhovitch, T. H. Nguyen, H. de Vries, and A. Courville. Systematic Generalization: What Is Required and Can It Be Learned? *International Conference on Learning Representations*, nov 2019. URL http://arxiv.org/abs/1811.12889.

[4] M. Baroni. Linguistic generalization and compositionality in modern artificial neural networks. mar 2019. URL http://arxiv.org/abs/1904.00157.

[5] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu. Relational inductive biases, deep learning, and graph networks. URL https://arxiv.org/pdf/1806.01261.pdf.

[6] Y. Bengio, N. Léonard, and A. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[7] B. Bogin, M. Geva, and J. Berant. Emergence of Communication in an Interactive World with Consistent Speakers. sep 2018. URL http://arxiv.org/abs/1809.00549.

[8] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[9] H. Brighton. Compositional syntax from cultural transmission. *MIT Press*, Artificial, 2002. URL https://www.mitpressjournals.org/doi/abs/10.1162/106454602753694756.

[10] H. Brighton and S. Kirby. The survival of the smallest: Stability conditions for the cultural evolution of compositional language. In *European Conference on Artificial Life*, pages 592–601. Springer, 2001.

[11] H. Brighton and S. Kirby. Understanding Linguistic Evolution by Visualizing the Emergence of Topographic Mappings. *Artificial Life*, 12(2):229–242, jan 2006. ISSN 1064-5462. doi: 10.1162/artl.2006.12.2.229. URL http://www.mitpressjournals.org/doi/10.1162/artl.2006.12.2.229.

[12] R. Chaabouni, E. Kharitonov, E. Dupoux, and M. Baroni. Anti-efficient encoding in emergent communication. *NeurIPS*, may 2019. URL http://arxiv.org/abs/1905.12561.

[13] R. Chaabouni, E. Kharitonov, A. Lazaric, E. Dupoux, and M. Baroni. Word-order biases in deep-agent emergent communication. may 2019. URL http://arxiv.org/abs/1905.12330.

[14] R. Chaabouni, E. Kharitonov, D. Bouchacourt, E. Dupoux, and M. Baroni. Compositionality and Generalization in Emergent Languages. apr 2020. URL http://arxiv.org/abs/2004.09124.

[15] E. Choi, A. Lazaridou, and N. de Freitas. Compositional Obverter Communication Learning From Raw Visual Input. apr 2018. URL http://arxiv.org/abs/1804.02341.

[16] F. Chollet. On the Measure of Intelligence. Technical report, 2019.

[17] N. Chomsky. *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group, 1986.

[18] K. Denamganaï and J. A. Walker. Referentialgym: A nomenclature and framework for language emergence & grounding in (visual) referential games. *4th NeurIPS Workshop on Emergent Communication*, 2020.

[19] J. A. Fodor, Z. W. Pylyshyn, et al. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[20] J. Foerster, I. A. Assael, N. de Freitas, and S. Whiteson. Learning to Communicate with Deep Multi-Agent Reinforcement Learning, 2016. URL http://papers.nips.cc/paper/6042-learning-to-communicate-with-deep-multi-agent-reinforcement-learning.

[21] S. Guo, Y. Ren, S. Havrylov, S. Frank, I. Titov, and K. Smith. The emergence of compositional languages for numeric concepts through iterated learning in neural agents. *arXiv preprint arXiv:1910.05291*, 2019.

[22] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

[23] S. Havrylov and I. Titov. Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols. may 2017. URL http://arxiv.org/abs/1705.11192.

[24] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, 2 (5):6, 2017.

[25] I. Higgins, N. Sonnerat, L. Matthey, A. Pal, C. P. Burgess, M. Botvinick, D. Hassabis, and A. Lerchner. SCAN: Learning Abstract Hierarchical Compositional Visual Concepts. jul 2017. URL http://arxiv.org/abs/1707.03389.

[26] F. Hill, A. Lampinen, R. Schneider, S. Clark, M. Botvinick, J. L. McClelland, and A. Santoro. Emergent Systematic Generalization in a Situated Agent. oct 2019. URL http://arxiv.org/abs/1910.00571.

[27] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[28] T.-W. Huang. Tensorboardx, 2018. URL https://github.com/lanpa/tensorboardX.

[29] D. Hupkes, V. Dankers, M. Mul, and E. Bruni. Compositionality decomposed: how do neural networks generalise? aug 2019. URL http://arxiv.org/abs/1908.08351.

[30] J. R. Hurford. Language and number: The emergence of a cognitive system. 1987.

[31] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

[32] Y. Jiang, S. Gu, K. Murphy, and C. Finn. Language as an Abstraction for Hierarchical Deep Reinforcement Learning. jun 2019. URL http://arxiv.org/abs/1906.07343.

[33] B. Keresztury and E. Bruni. Compositional properties of emergent languages in deep learning. jan 2020. URL http://arxiv.org/abs/2001.08618.

[34] S. Kirby. Learning, bottlenecks and the evolution of recursive syntax. 2002.

[35] S. Kirby and J. R. Hurford. The Emergence of Linguistic Structure: An Overview of the Iterated Learning Model. In *Simulating the Evolution of Language*, pages 121–147. Springer London, 2002. doi: 10.1007/978-1-4471-0663-0_6.

[36] S. Kirby, T. Griffiths, and K. Smith. Iterated learning and the evolution of language. *Current opinion in neurobiology*, 28:108–114, 2014. URL https://www.sciencedirect.com/science/article/pii/S0959438814001421.

[37] T. Korbak, J. Zubek, Ł. Kuciński, P. Miłoś, and J. Rączaszek-Leonardi. Developmentally motivated emergence of compositional communication via template transfer. oct 2019. URL http://arxiv.org/abs/1910.06079.

[38] K. Korrel, D. Hupkes, V. Dankers, and E. Bruni. Transcoding compositionally: using attention to find more generalizable solutions. jun 2019. URL http://arxiv.org/abs/1906.01234.

[39] S. Kottur, J. M. F. Moura, S. Lee, and D. Batra. Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog. jun 2017. URL http://arxiv.org/abs/1706.08502.

[40] M. Krifka. Compositionality. *The MIT encyclopedia of the cognitive sciences*, pages 152–153, 2001.

[41] B. M. Lake. Compositional generalization through meta sequence-to-sequence learning. jun 2019. URL http://arxiv.org/abs/1906.05381.

[42] B. M. Lake and M. Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *35th International Conference on Machine Learning, ICML 2018*, 7:4487–4499, oct 2018. URL http://arxiv.org/abs/1711.00350.

[43] A. Lazaridou, A. Peysakhovich, and M. Baroni. Multi-Agent Cooperation and the Emergence of (Natural) Language. dec 2016. URL http://arxiv.org/abs/1612.07182.

[44] A. Lazaridou, K. M. Hermann, K. Tuyls, and S. Clark. Emergence of Linguistic Communication from Referential Games with Symbolic and Pixel Input. apr 2018. URL http://arxiv.org/abs/1804.03984.

[45] F. Li and M. Bowling. Ease-of-Teaching and Language Structure from Emergent Communication. jun 2019. URL http://arxiv.org/abs/1906.02403.

[46] A. Liška, G. Kruszewski, and M. Baroni. Memorize or generalize? Searching for a compositional RNN in a haystack. feb 2018. URL http://arxiv.org/abs/1802.06467.

[47] J. Loula, M. Baroni, and B. M. Lake. Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks. jul 2018. URL http://arxiv.org/abs/1807.07545.

[48] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[49] C. J. Maddison, A. Mnih, and Y. W. Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *International Conference on Learning Representations*, 2017.

[50] G. Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

[51] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[52] J. L. McClelland, M. M. Botvinick, D. C. Noelle, D. C. Plaut, T. T. Rogers, M. S. Seidenberg, and L. B. Smith. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, 14(8):348–356, 2010.

[53] R. Montague. Universal grammar. *Theoria*, 36(3):373–398, 1970.

[54] I. Mordatch and P. Abbeel. Emergence of Grounded Compositional Language in Multi-Agent Populations. URL https://arxiv.org/pdf/1703.04908.pdf.

[55] K. Moritz Hermann, F. Hill, S. Green, F. Wang, R. Faulkner, H. Soyer, D. Szepesvari, W. M. Czarnecki, M. Jaderberg, D. Teplyashin, M. Wainwright, C. Apps, D. Hassabis, P. Blunsom, and D. London. Grounded Language Learning in a Simulated 3D World. URL https://arxiv.org/pdf/1706.06551.pdf.

[56] T. pandas development team. pandas-dev/pandas: Pandas, Feb. 2020. URL https://doi.org/10.5281/zenodo.3509134.

[57] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL `http://jmlr.org/papers/v12/pedregosa11a.html`.

[59] F. Perez and B. E. Granger. Ipython: A system for interactive scientific computing. *Computing in Science Engineering*, 9(3):21–29, 2007. doi: 10.1109/MCSE.2007.53.

[60] Y. Ren, S. Guo, M. Labeau, S. B. Cohen, and S. Kirby. Compositional Languages Emerge in a Neural Iterated Learning Model. feb 2020. URL `http://arxiv.org/abs/2002.01365`.

[61] C. Resnick, A. Gupta, J. Foerster, A. M. Dai, and K. Cho. Capacity, Bandwidth, and Compositionality in Emergent Language Learning. oct 2019. URL `http://arxiv.org/abs/1910.11424`.

[62] J. Russin, J. Jo, R. C. O'Reilly, and Y. Bengio. Compositional generalization in a deep seq2seq model by separating syntax and semantics. apr 2019. URL `http://arxiv.org/abs/1904.09708`.

[63] A. Słowik, A. Gupta, W. L. Hamilton, M. Jamnik, S. B. Holden, and C. Pal. Exploring Structural Inductive Biases in Emergent Communication. feb 2020. URL `http://arxiv.org/abs/2002.01335`.

[64] K. Smith, S. Kirby, H. B. A. Life, and U. 2003. Iterated learning: A framework for the emergence of language. *Artificial Life*, 9(4):371–389, 2003. URL `https://www.mitpressjournals.org/doi/abs/10.1162/106454603322694825`.

[65] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[66] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

[67] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[68] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

[69] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

# A Supplementary Materials

Figure 4 (top two rows) illustrates the impact on both the topographic similarity and the accuracy of the zero-shot compositional tests when varying the maximum sentence length $L$, while keeping the vocabulary size $V$ fixed, and vice versa (bottom two rows), as discussed in Section 4.2.2.

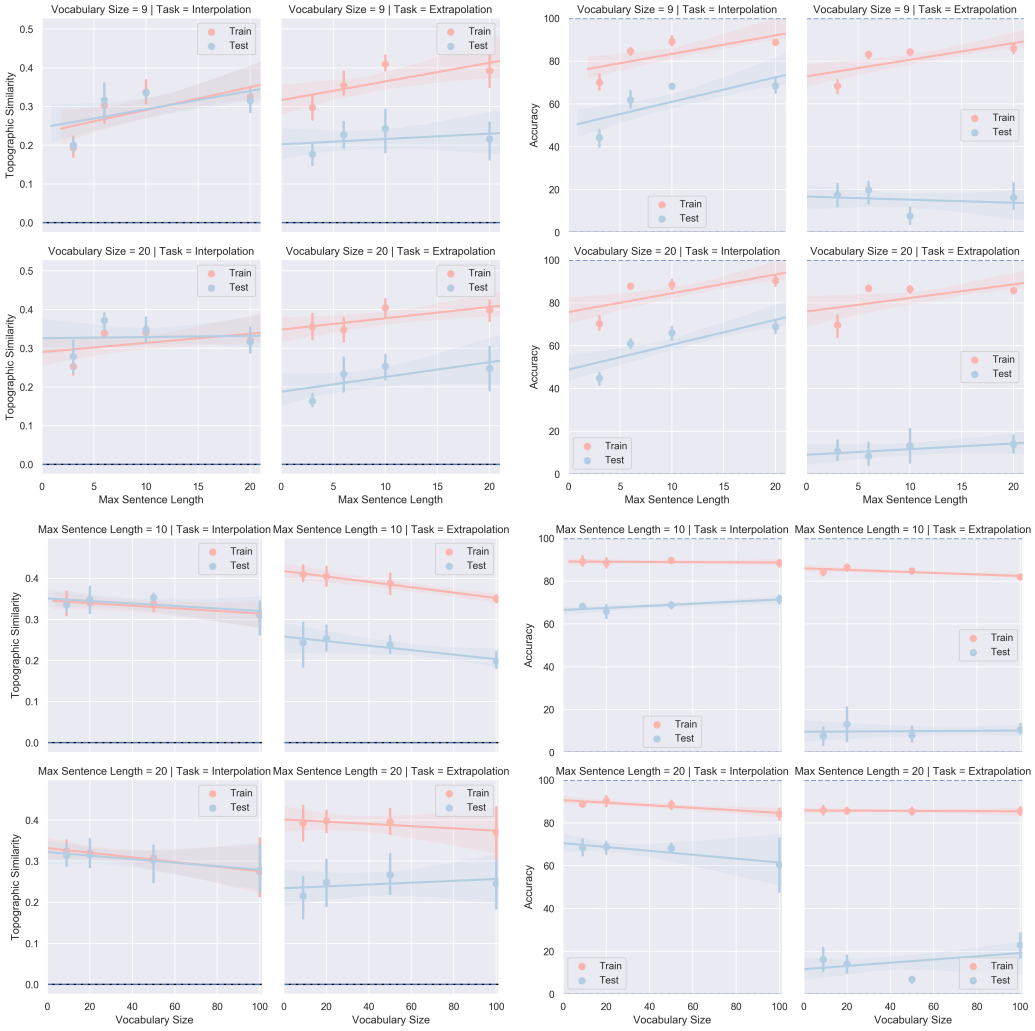


Figure 4: Distributions of the topographic similarity and accuracy, across different traversals of the communication channels capacities, on the 3-attributes benchmarks. **Top**: traversal of the max sentence length $L \in \{3, 6, 10, 20\}$, for fixed $V \in \{9, 20\}$. **Bottom**: traversal of the vocabulary size $V \in \{10, 20, 50, 100\}$, for fixed $L \in \{10, 20\}$.

# B Language Emergence

Compositionality and recursive syntax are the two main properties shared between all natural languages that account for their expressivity and flexibility. Among other things, they allow an ease of learning/acquisition as it can be seen in the *poverty of the stimulus* phenomenon ("children master complex features of language on the basis of surprisingly little evidence" [9]).

When talking about language, it is common to differentiate between two language domains: I(Internal)-Language and E(External)-Language [17, 30, 34] . I-Language is the internal representation a user has of the language it speaks and listens to, whilst E-Language is the set of external presentations/"utterances in the arena of use"[30] of language, when any of the users actually speak in an attempt to express a given meaning.

How language users are able to acquire an I-Language and are then able to contribute to the E-Language by producing utterances of language that exhibit compositionality is the question that the works of [34, 10, 9, 64] strived to answer. To achieve this they appealed to a "process of information transmission via observational learning" [9], which took place in-between populations and generations of (previously-learner) speaker agents and (soon-to-become-speaker) learner agents. This is the Iterated Learning Model (ILM) [36].

In this section, we will compare the ST-GS estimator to the ILM in order to highlight insights about the drivers of compositional language emergence using the ST-GS estimator.

## B.1 (Neural) Iterated Learning Model

The ILM consists of a "[multi-]agent-based model where each agent represents a language user"[9], which forms utterances based on the hypothesis that has been formed previously (I-Language) to account for previously-observed utterances (E-Language). It puts in relation an environment, made up of $N$ objects/stimuli that maps to a meaning space $\mathcal{M} = \{(f_i)_{i\in[1:F]} \ / \ \forall f_i, 1 \leq f_i \leq V\}$, (consisting of an $F$-dimensional space, i.e. features, with $V$ possible discrete values along each dimension) with a signal space, or language, $\mathcal{S} = \{(\omega_i)_{i\in[1:l]} \ / \ \forall \omega_i \in \Sigma \ , \ \forall l \in [1 : l_{max}]\}$ (a set of strings of symbols from a symbol space $\Sigma$ with length $l \leq l_{max}$). The relation or mapping is induced by making agents observe (attend to) signal-meaning pairs from the E-Language, during the *acquisition phase*, and then, during the *production phase*, making them utter new language utterances, whilst being prompted to communicate/speak about a set of other randomly-sampled objects/stimuli from the environment.

### B.1.1 Language Stability

In such a context where language is evolved by each learner agent at each iteration, it becomes important to consider the stability of the different languages that may arise, in order to understand under which conditions does the ILM converge and what are the properties to expect of the emerging, stable languages. Before being able to consider the stability of a given language, we need to define its expressivity $E$, especially "the expected number of meanings an individual will be able to express after observing some subset of the E-Language"[64], which we will denote $O$.

**Holistic Language** - In the context of a holistic language where there is no syntactic structure to the language and it consists of idiosyncrasies, no *generalisation* can be operated, only *memorisation* upon observation and then recall can help an agent express a given meaning in that language. Therefore, the expressivity of a holistic language $E_h$ "is simply the probability of observing any particular meaning [$m \in O$, $Pr(m \in O)$,] multiplied by the number of possible meanings" [64], as shown in Eqn. 1.

$$E_h = Pr(m \in O) \cdot \binom{V}{1}^F = Pr(m \in O) \cdot V^F. \tag{1}$$

It is worth noting that the meaning space $\mathcal{M}$ is structured as an $F$-dimensional space of features with $V$ possible values for each feature.

**Compositional Language** - In the context of a compositional language where the relationship between meanings and signals is structured, "expressivity becomes a function of the number of feature values observed, rather than a function of the number of meanings observed" [9]. Given a meaning $m = (v_1, \ldots, v_F)$, formally, we have:

$$E_c = Pr(\text{“}\forall v \in m, \exists O_v \in R, v \in O_v\text{”}) \cdot N_{used} \tag{2}$$

where $Pr(\text{“}\forall v \in m, \exists O_v \in R, v \in O_v\text{”}) = Pr(\text{“}\exists O_v \in R, v \in O_v\text{”})^F$ is the probability of being able to express $m$ (to be contrasted with the probability of observing $m \in O$, $Pr(m \in O)$), and thus, $N_{used}$ is “the expected number of expressible meanings" [9] that are used to label the $N$ objects in the environment. The more eager readers can refer to [9] for more details.

#### B.1.2 Drivers of Compositionality

Since we are concerned with the likelihood of the emergence of compositionality over holisticity, the authors define in Eqn. 3 the relative stability $S$ of compositional languages with respect to holistic ones:

$$S = \frac{S_c}{S_c + S_h}, \tag{3}$$

where $S_c \propto \frac{E_c}{N}$ and $S_h \propto \frac{E_h}{N}$ are the (absolute) stabilities of compositional languages and holistic ones, respectively. Thus, “the probability of observing some arbitrary meaning $m$, i.e. $Pr(m \in O)$, is determined by the number of objects in the environment ($N$), the number of meanings in the meaning space ($M$, where $M = V^F$), and the number of random object observations during an agent lifetime ($R$)" [9]. Therefore, the relative stability of a compositional language compared to a holistic language will be strongly affected by: (i) the structure of the meaning space, via its dependence to $M$, and (ii) the object coverage expressed by the ratio $b = \frac{R}{N}$ (also the inverse of the measure of the severity of the *transmission bottleneck*). The lower the coverage, i.e. $b$, the greater the stability advantage of compositional languages over holistic ones is, such that “the poverty-of-the-stimulus ‘problem’ is in fact required for linguistic structure to emerge” [64]. Another important result is visible in the observation that “a large stability advantage for a compositional language (high $S$) only occurs when the meaning space exhibits a certain degree of structure (i.e. when there are many features and/or values), suggesting that structure in the conceptual space of language learners is a requirement for the evolution of compositionality" [64].

In other words, both the severity of the (cultural) *transmission bottleneck* and the degree of structure in the meaning space have been identified as drivers of compositionality in emerging languages, in the context of the ILM.

Recently, the Neural Iterated Learning (NIL) algorithm was proposed by Ren et al. [60]. They transposed the ILM framework to a deep learning setting and showed a similar impact of the severity of the transmission bottleneck, in addition to a proposed probabilistic framework that formally evidenced it.

### B.2 Straight-Through Gumbel-Softmax Estimator

In the current computer science paradigm, we commonly assume a discrete nature of the messages that are sent by the *speaker* to the *listener*. Therefore, the communication channel is non-differentiable and must rely on *Reinforcement Learning* algorithms to solve the credit-assignment problem. The most common candidates in the literature are REINFORCE-like algorithms [69]. Fortunately, tricks exist that allow the environment (in this case, the communication channel) to be made differentiable.

As it takes place in the context of an $L \geq 2$-*signal*/0-*round* referential game, the work of Havrylov and Titov [23] brings evidence that deep learning agents can not only learn to coordinate via a communication channel (as seen in Lazaridou et al. [43] with “atomic symbols” [23] already), but first and foremost invent a seemingly compositional, hierarchical, and variable-length language, with a “sequence of tokens” [23], in order to coordinate. The model’s greatest success is attributed to the introduction of the Straight-Through Gumbel-Softmax (ST-GS) estimator/relaxation/backpropagation approach, which makes the communication channel differentiable (and also exhibits similar behaviour between training and testing, which is on the contrary to previous works, e.g. Foerster et al. [20]).

The ST-GS estimator is built on top of the Gumbel-Softmax estimator [31, 49], that replaced one-hot-encoded symbols/tokens/words $w \in V$, originally sampled from a categorical distribution, with a continuous relaxation $\tilde{w}$, sampled from a Gumbel-Softmax distribution, following the notation of Havrylov and Titov [23]:

$$\tilde{w_k} = \frac{\exp((\log p_k + g_k)/\tau)}{\sum_{i=1}^{K} \exp((\log p_i + g_i)/\tau)}$$

where $p_1, \ldots, p_K$ are the $K$ event probabilities of the original categorical distribution, $\tau$ is a temperature hyperparameter (see [23] for more details), and $g_1, \ldots, g_K$ are sampled from the Gumbel distribution, i.e. $g_k = -\log(-\log(u_k))$ with $u_k \sim \mathcal{U}(0, 1)$.

Instead of stopping there, the ST-GS estimator performs a greedy discretization (i.e. using the $argmax$ operator) of $\tilde{w_k}$ during the forward pass, whilst relying on the continuous relaxation during the backward pass[2], thus yielding a biased gradient estimator [31, 6].

Havrylov and Titov [23] simplified the whole process by considering learning the temperature $\tau(h_i^s)$ for each symbol/token/word $w_i$ in each sentence $s$ with conditioning on the hidden state $h_i^s$ of the sentence decoder RNN. In this work, instead of using a multi-layer perceptron, we only rely on a one-layer network $\alpha$:

$$\tau(h_i^s) = \frac{1}{\tau_0 + \log(1 + \exp(\alpha(h_i^s)))}$$

One very important and somewhat surprising result that they found is that using the ST-GS estimator, the greater the sequence length, the faster the learning of the communication protocol. On the other hand, there is no such correlation when training REINFORCE-like algorithms. It is unclear whether this phenomena is due to the ST-GS estimator alone or the synergy between it and the pre-trained convolutional neural network (CNN), which each agent relied on in the original work. In order to efficiently assess the impact of the ST-GS estimator, in this paper we chose to have our agents learn everything from scratch.

### B.3 Instantiating a Transmission Bottleneck

In the following, we highlight the similarities between the ILM, in the form of the NIL algorithm, and the ST-GS algorithm. Both the NIL and the ST-GS algorithms are iterative processes. Thus, we focus on one learning step for each.

Firstly, in the ILM/NIL, the *learning phase* consists of the new learning agent $A_i$ updating its model using the data set $\mathcal{D}_{i-1} = (u_j^{i-1}, s_j)_{j \in [1; N_{learning}]}$ of pairs of stimuli $s$ and associated utterance $u$ produced by agent $A'_{i-1}$ at the previous iteration. $N_{learning}$ is the size of the learning data set, at each learning step, and it is randomly sampled from the whole meaning space $\mathcal{M}$ of size $N$. In the ILM/NIL, a cultural transmission bottleneck is instantiated by choosing $N_{learning} \leq N$, and the severity of the bottleneck is measured by $R = \frac{N_{learning}}{N}$. Next, following the naming of Ren et al. [60], in the *transmitting phase*, the now updated agent $A'_i$, fluent in the E-language $\mathcal{L}_{i-1}$ (as described by $\mathcal{D}_{i-1}$), is prompted to a production step where it generates the data set $\mathcal{D}_i = (A'_i(\hat{s}_j), \hat{s}_j)_{j \in [1; N_{learning}]}$ with stimuli being randomly sampled, $\hat{s} \sim \mathcal{M}$. Due to the transmission bottleneck during the production step, the learning agent has to generalise its knowledge of the E-language $\mathcal{L}_{i-1}$ to potentially novel stimuli, and it is thus bound to make this language evolve into a new E-language $\mathcal{L}_i$, unless there is no transmission bottleneck and it has perfectly learned the language during the learning step. In the NIL algorithm, an *interacting phase* is intertwined between the *learning phase* and the *transmitting phase*, which consists of a referential game $\mathcal{R}$ that aims to promote disambiguation of the language.

It is important to note that the ILM/NIL algorithm relies on the previous E-language $\mathcal{L}_{i-1}$ at each step $i$, whilst the ST-GS algorithm relies on something more akin to the previous I-language, as we will now detail.

In the context of the ST-GS algorithm, learning happens by batch. At each learning step, a batch $\mathcal{D}_i^S$ of stimuli, randomly sampled from the whole meaning space $\mathcal{M}$ (as defined in Eqn. 4), is forwarded through the learning *speaker* agent $S_i$ to produce a batch $\mathcal{D}_i^L$ of utterances (as defined in Eqn. 5). The latter is then forwarded through the learning *listener* agent $L_i$, along with a batch of sets of stimuli $\Delta_i$. Each element of the batch is a shuffled list/set containing $K$ distractor stimuli that are randomly sampled from $\mathcal{M}$, and the corresponding target stimuli from $\mathcal{D}_i^S$ (as defined in Eqn. 6). The resulting output is a batch of sets of scores $\Sigma_i$ (as defined in Eqn. 7). Using $j$ as the index for the element of the batch, each score $s(u_j^i, \hat{(s)}_j^i(d)) \in \sigma_i^j$ intuitively represents the extent with which each utterance $u_j^i$ describes the stimuli $\hat{(s)}_j^i(d) \in \delta_i^j$.

[2] for further implementation details see: `https://pytorch.org/docs/master/nn.functional.html?highlight=gumbel#torch.nn.functional.gumbel_softmax`

$$\mathcal{D}_i^S = (s_j^i)_{j \in [1; N^S_{learning}]} \sim \mathcal{M} \tag{4}$$

$$\mathcal{D}_i^L = (u_j^i = S_i(s_j^i))_{j \in [1; N^S_{learning}]} \tag{5}$$

$$\Delta_i = (\delta_i^j)_{j \in [1; N^S_{learning}]}, s.t. \quad \forall j \in [1; N^S_{learning}], \delta_i^j = (\hat{s}_j^i(d))_{d \in [1; K+1]} \tag{6}$$

$$\Sigma_i = (\sigma_i^j)_{j \in [1; N^S_{learning}]} \tag{7}$$

Following the computation of the loss function, similarly to Havrylov and Titov [23], the gradients are backpropagated and the agents are updated. It is important to notice that at the time of the backward pass, the current agents are not yet updated, i.e. $L_i = L_{i-1}$ and $S_i = S_{i-1}$. Therefore, each of them receives a gradient with respect to the other's I-language (as represented by each agent's weights) at the previous time step, i.e. with respect to $S(\mathcal{L}_{i-1})$ and $L(\mathcal{L}_{i-1})$ respectively. Following the backward pass, the algorithm yields the updated agents $S_i'$ and $L_i'$.

In comparison to the NIL algorithm, we argue that both updated agents have received feedback in a richer fashion due to: (i) learning in a supervised fashion (that not only promotes but also penalises the E-language with respect to the goal of disambiguation) and (ii) the fact that the ST-GS relaxation enables feedback with respect to both the referential game $\mathcal{R}$ and the learning *listener* agent's I-language $L(\mathcal{L}_{i-1})$ (which we assume to be more pertinent than the E-language $\mathcal{L}_{i-1}$).

# C Agent Architecture

Each agent consist of a language module and a visual or symbolic module, depending whether they deal with visual or symbolic stimuli. The *listener* agent also incorporates a third decision module that combines the outputs of the other two (visual and language) modules. While the *speaker* agent is prompted to produce the output string of symbols with a *Start-of-Sentence* symbol and the visual module output as an initial hidden state, the *listener* agent consumes the string of symbols with the null vector as the initial hidden state. In the following subsections, we detail each module architecture in depth.

## C.1 Visual & Modules

The visual module $f(\cdot)$ consists of four $3 \times 3$ convolutional layers with stride 2. The two first layers have 32 filters, whilst the last two layers have 64. Each convolutional layer is followed by a 2D batch normalisation layer, and the outputs are passed through a leaky ReLU activation function. Inputs are resized to $32 \times 32$, thus yielding feature maps of dimension $64 \times 2 \times 2$. The visual module outputs a flattened representation of dimension 256. In a similar fashion, the symbolic module consists of a linear layer containing 256 units followed by a leaky ReLU activation function.

## C.2 Language Module

The language module $g(\cdot)$ consists of a one-layer LSTM network [27] with 256 hidden units, matching the dimension of the visual module output. In the context of the *listener* agent, the input message $m = (m_i)_{i \in [1,L]}$ (produced by the *speaker* agent) is represented as a string of one-hot encoded vectors of dimension $|V|$ and embedded in an embedding space of dimension 256 via a linear layer and dropout layer of probability $p = 0.8$. The output of the *listener* agent's language module, $g^l(\cdot)$, is the last hidden state of the LSTM layer, $h^l_L = g^l(m_L, h^l_{L-1})$. In the context of the *speaker* agent's language module $g^s(\cdot)$, the output is the message $m = (m_i)_{i \in [1,L]}$ consisting of one-hot encoded vectors of dimension $|V|$, which are sampled using the ST-GS approach from a categorical distribution $Cat(p_i)$ where $p_i = Softmax(\nu(h^s_i))$, provided $\nu$ is an affine transformation and $h^s_i = g^s(m_{i-1}, h^s_{i-1})$. $h^s_0 = f(s_t)$ is the output of the visual module, given the target stimulus $s_t$.

## C.3 Decision Module

Similar to Havrylov and Titov [23], the decision module builds a probability distribution over a set of $K + 1$ stimuli/images $(s_0, ..., s_K)$, consisting of the target stimulus and $K$ distractor stimuli, given a message $m$:

$$p((s_i)_{i \in [0,K]}|m) = Softmax((h^l_L \cdot f(s_i)^T)_{i \in [0,K]}).$$

# D Lack of Correlation between Generalisation Abilities and Training-time Topographic Similarity

Section 4.2.3 provided evidence that compositionality in the E-language at training-time, as reported by topographic similarity, cannot be expected to correlate with the generalisation abilities of the learning agents, without differentiating between the different settings. In this section, we provide a closer look at the results for each benchmark. Figure 5 shows the test-time accuracy with respect to the training-time topographic similarity for each different setting. The results of conducting Spearman rank-order correlation tests on the data yields correlation coefficients as described in Table 2.

Strikingly, the results emphasise the difference between the interpolation task and the extrapolation task. While the results in the case of the interpolation task hint at a positive correlation hypothesis (the lower the level of structure in the observed meaning space is), those in the extrapolation task tend towards a negative correlation, for its most significant result (the higher the level of structure is). Our results are diverse and dependant on the level of structure in the observed meaning space, in addition to the type of task. This breakdown analysis further confirms our previous results that prevents us from rejecting the non-correlation hypothesis, and piles up further evidence to show that it would be incorrect to expect the training-time topographic similarity to predict any generalisation abilities of the learning agents.

# E Generalisation Difficulty in Interpolation & Extrapolation

**Generalisation difficulty**, as defined by Chollet [16], can be intuitively understood as a measure of how much the optimal behaviour of a skill program (here represented by the learning agent's weights) at evaluation-time differs from the optimal behaviour at training-time. The skill program is the product of an intelligent system (here represented by the whole optimisation process and surrounding algorithms that update the deep learning agents' weights), that we seek to evaluate the intelligence of. It is formally defined in an Algorithmic Information Theory fashion (considering program in terms of their description string), for a given task, $T$ (consisting of training and testing phases), a given curriculum, $C$ (here represented by the training scheme employed, i.e. a discriminative referential game), and a given skill threshold, $\theta$, as the "length of the shortest program that, taking as input the shortest possible program that performs optimally over the situations in curriculum $C$, produces a program that performs at a skill level of at least $\theta$ during evaluation, normalized by the length of that skill program" [16]. In other words, the (Relative) Algorithmic/Kolmogorov Complexity: $GD^{\theta}_{T,C} = H(Sol^{\theta}_{T}|TrainSol^{opt}_{T,C})/H(Sol^{\theta}_{T})$ where $Sol^{\theta}$ is "the shortest of all possible solutions to $T$ of threshold $\theta$ (shortest skill program that achieves at least skill $\theta$ during evaluation)", and $TrainSol^{opt}_{T,C}$ is "the shortest optimal training-time solution given a curriculum (shortest skill program that achieves optimal training-time performance over the [stimuli] in the curriculum)". Therefore, $GD^{\theta}_{T,C}$ is the normalised description length (in a fixed universal language) of the shortest intelligent program that outputs $Sol^{\theta}_{T}$ when taking as input $TrainSol^{opt}_{T,C}$.

Table 2: Results of Spearman rank-order correlation tests for each benchmark, independent of the communication channel capacity (i.e. 5 seeds in the complete case, and 5 other seeds in the overcomplete case, for each benchmark).

| Benchmark | Correlation Coefficient | P-value |
|---|---|---|
| $T^2_{inter}$ | 0.811 | 0.004 |
| $T^3_{inter}$ | 0.601 | 0.066 |
| $T^4_{inter}$ | 0.115 | 0.750 |
| $T^2_{extra}$ | 0.356 | 0.313 |
| $T^3_{extra}$ | 0.370 | 0.294 |
| $T^4_{extra}$ | −0.503 | 0.138 |

### E.1 Experiments

Following our primer on the definition of generalisation difficulty, in Section E, we propose to investigate the generalisation difficulty of the interpolation task $T_{inter}$ and extrapolation task $T_{extra}$, for a skill threshold $\theta = 38\%$ and our curriculum $C$ being a discriminative referential game with $K = 47$ distractors and overcomplete communication channel. Defining optimality on the training phase of each task as achieving performance of at least $80\%$ accuracy, we find that all our different seeds in $T^3_{inter}$ and $T^3_{extra}$ can be considered optimal training-time solutions. We assume that any description of the learning agents are the shortest possible among the viable deep learning agents that can undertake our tasks. Our learning agents have minimal architecture design for language emergence and grounding (see Appendix C for details). It ensues that the identity function or program is a viable program to transform our distribution of shortest optimal training-time skill programs $TrainSol^{opt}_{T^3_{inter},C}$ into skill programs that perform at a skill level of $\theta = 38\%$ on the 3-attributes interpolation task $T^3_{inter}$. We acknowledge that the identity program is the shortest description-length program possible on any fixed universal language, therefore it is the description-length of the shortest intelligent program that outputs $Sol^{\theta=38\%}_{T^3_{inter}}$ when taking as input $TrainSol^{opt}_{T^3_{inter},C}$.

On the other hand, for the extrapolation task $T^3_{extra}$, Figure 6 shows that the identity program is not enough to transform the input $TrainSol^{opt}_{T^3_{extra},C}$ into $Sol^{\theta=38\%}_{T^3_{extra}}$, as the distributions of skill programs that are the shortest optimal training-time solutions on $T^3_{extra}$ perform below the skill level $\theta = 38\%$. We interpret this result as the fact that the shortest description-length program transforming $TrainSol^{opt}_{T^3_{extra},C}$ into $Sol^{\theta=38\%}_{T^3_{extra}}$ must be more complex than the identity program (i.e. of a longer description-length). It ensues that $GD^{\theta=38\%}_{T^3_{extra},C} > GD^{\theta=38\%}_{T^3_{inter},C}$ (i.e. given the curriculum $C$ and skill threshold $\theta = 38\%$, the generalisation difficulty of task $T^3_{extra}$ is higher than that of task $T^3_{inter}$).

From a different and less formal viewpoint, Figure 6 shows the test-train gap in terms of topographic similarity of the E-language, across different settings. Independent of the level of structure in the meaning space, or the capacity of the communication channel, the topographic similarity test-train gap is consistently and significantly worse in the $T_{extra}$ ($p \approx 4e^{-10}$ for a one-sided Kolmogorov-Smirnov test with the alternative hypothesis being that the distribution of topographic similarity test-train gap in $T_{extra}$is lower than the one of $T_{inter}$), showing that the speaker agent fails to generalise systematically as the produced utterances at testing time are fairly different than at training time. We argue that this difference of behaviour that exists in $T_{extra}$ and not in $T_{inter}$ (the median of the distribution is at 0), shows the extent with which the latter requires the behaviour of the agents to differ from the training time to the evaluation time. As this is another intuition about generalisation difficulty [16], our results lend themselves to the conclusion that indeed there is some generalisation difficulty in $T_{extra}$ (in the order of at least $18\%$, as the median of the test-train gap in topographic similarity in $T_{extra}$), whilst there is none in $T_{inter}$.

We expect this result to hold across curriculum for the family of visual referential games, independently of the technique that supports the communication channel, and for tasks where the stimuli are produced by a generative process, which take attribute vectors as input. For instance, referential games with symbolic stimuli are not expected to abide by these results.

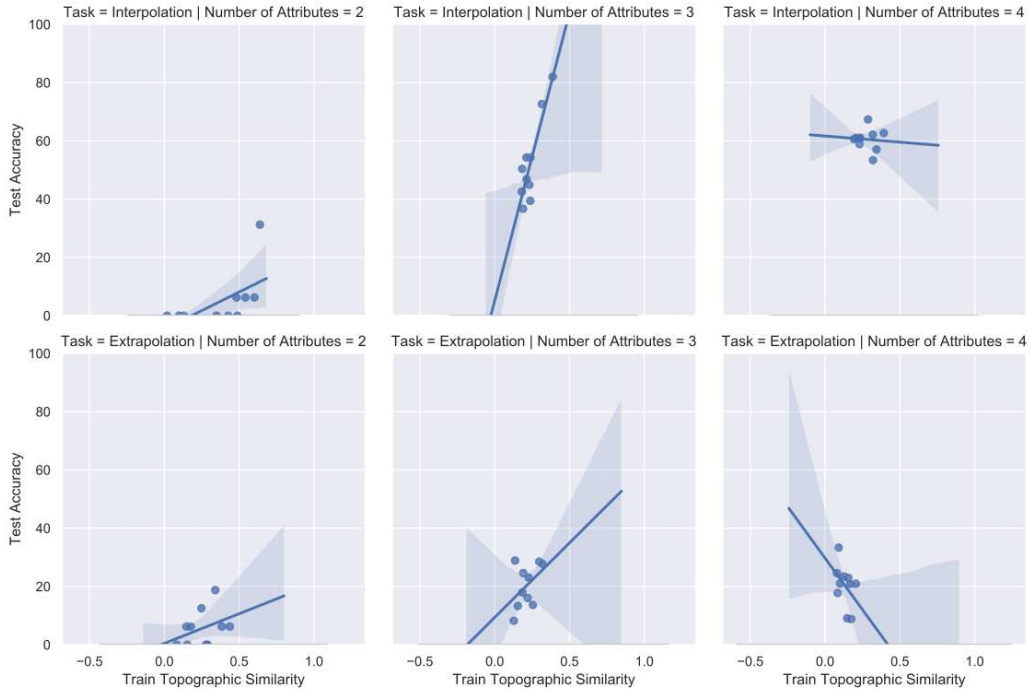


Figure 5: Training-time topographic similarity Vs. test-time accuracy for each benchmark. A linear regression model is fit to the data for comparison.

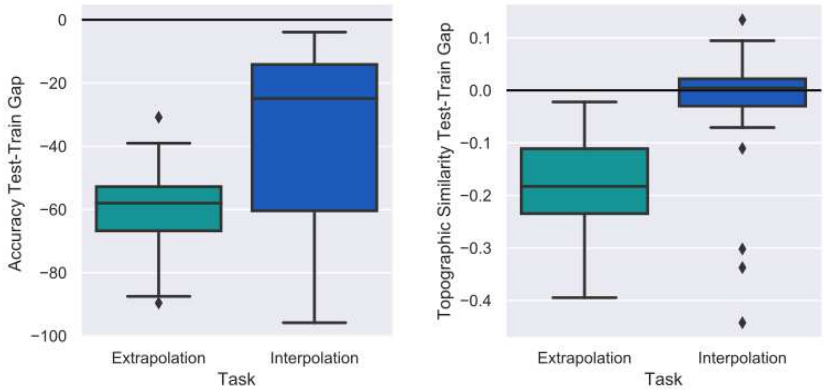


Figure 6: Distributions of the test-train gap in terms of accuracy and topographic similarity of emerging languages in different settings.